**IQVIA pharma records anomalies detection**

**IQVIA data architect**: Angel Rey

**Presentation Title:** Automatic anomalies detection on pharma records

**Research Focus:** Distributed, scalable machine learning applied to pharma dataset

**Company Department**: Data acquisition, Big Data Factory

**Presentation Type**: Running demo

**Abstract**:

Machine Learning pipeline proposed as a PoC will consists in a distributed not supervised clustering algorithms plus ml techniques to clean, normalize and convert data and reduce dimensions (if required) as previous steps providing as result candidate records suspected of being anomalous for further inspection.

Also, as part of the process other ml calculations like the selection of the main parameters of the algorithm will be developed as well.

Approach will use Spark and Scala with k-means as initial algorithm. The concept of “anomaly” it will be defined as a point that it is assigned to a cluster, but it is too far from the centroid of the cluster that it belongs. The concept of “far” it is a parameter that can be adjustable by the user. The initial distance will be the Euclidian one but as part of future extension of the PoC different concept of anomaly, algorithm and distance could be chosen.

Different methods to choose k it will be used: Elbow method, Average Silhouette method and Gap statistic method, low entropy method.

Related to the feature normalization standard formula will be used: (feature\_i – μi) / σi

The conversion of categorical variables into numeric features will be done using one-hot encoding.

If required, random matrices, PCA or Singular Value Decomposition will be applied in order to reduce dimensions.

**Extension**: Migrate to a Realtime version using streaming and Kafka. Add other algorithms like Gaussian mixture model or DBSCAN. Add other distances like Minkowski distance or Manhattan distance.

**High level diagram:**

